WHAT IS BIG DATA?

Sources, Characteristics, Use Case and Hadoop-Based Solution

Presented by: Rachit Joshi

1. Introduction

Big Data refers to data that is very large in size. TPointTech explains that while everyday data may be measured in megabytes or gigabytes, data in petabytes (10^15 bytes) is described as Big Data. The page also states that almost 90% of today's data had been generated during the previous three years.

2. What is Big Data?

Big Data is data whose size is extremely large compared with the data normally handled in everyday applications. TPointTech uses the difference between megabyte- or gigabyte-sized data and petabyte-sized data to explain the concept.

The challenge with Big Data is not only its size. Large volumes of data can come from many different sources and may include structured as well as unstructured information. Such data needs suitable storage, processing, and analysis methods.

3. Sources of Big Data

TPointTech identifies several important sources that generate large amounts of data:

- Social Networking Sites – Platforms such as Facebook, Google, and LinkedIn generate large amounts of data because of their huge user bases.
- E-commerce Sites – Websites such as Amazon, Flipkart, and Alibaba generate logs that can be used to understand users' buying trends.
- Weather Stations – Weather stations and satellites generate large amounts of data that can be stored and manipulated for weather forecasting.
- Telecom Companies – Telecom companies study user trends and store data generated by millions of users.
- Share Market – Stock exchanges around the world generate large amounts of data through daily transactions.

These examples are directly listed on TPointTech's “What is Big Data” page.

4. 3 V’s of Big Data

V	Meaning	Description
Velocity	Speed of data generation	Data is increasing at a very fast rate; TPointTech states that the volume of data is estimated to double every two years.
Variety	Different forms of data	Data may be structured or unstructured. Log files and CCTV footage are examples of unstructured data, while tabular bank transaction data is structured.
Volume	Amount of data	The amount of data being handled can be extremely large, such as petabytes of information.

The 3 V's presented by TPointTech are Velocity, Variety, and Volume.

5. Velocity

Velocity describes how quickly data is being generated or increased. TPointTech states that data is increasing at a very fast rate and gives an estimate that the volume of data will double every two years.

6. Variety

Variety refers to the different forms in which data exists. Modern data is not limited to rows and columns. TPointTech distinguishes between structured data, such as bank transaction data stored in tables, and unstructured data, such as log files and CCTV footage.

7. Volume

Volume refers to the enormous amount of data that organizations and systems need to store and process. TPointTech uses petabytes as an example of the scale associated with Big Data.

8. Big Data Use Case

TPointTech presents an e-commerce example to show how Big Data can be used in practice. Suppose an e-commerce company named XYZ has 100 million users and wants to give a $100 gift voucher to its top 10 customers based on their spending during the previous year.

The company also wants to identify the buying trends of those customers so that it can recommend more relevant products to them. This requires the company to process and analyze a very large amount of customer and transaction data.

9. Issue in the Use Case

The main issue identified by TPointTech is the huge amount of unstructured data that needs to be stored, processed, and analyzed. A conventional approach may not be suitable for handling such a large dataset efficiently.

10. Solution for Big Data

TPointTech describes a Hadoop-based solution with separate approaches for storage, processing, analysis, and cost:

- Storage – Hadoop uses HDFS (Hadoop Distributed File System). HDFS uses commodity hardware to form clusters and stores data in a distributed fashion. The page describes its approach as a write-once, read-many-times principle.
- Processing – The MapReduce paradigm is applied to data distributed across the network to obtain the required output.
- Analysis – Tools such as Pig and Hive can be used to analyze the data.
- Cost – Hadoop is open source, so the cost of the software is not considered an issue in the presented solution.

These four parts—Storage, Processing, Analyze, and Cost—are the solution presented by TPointTech for the example.

11. HDFS for Big Data Storage

HDFS stands for Hadoop Distributed File System. In the TPointTech example, HDFS is used to store the huge amount of data by distributing it across a cluster of commodity hardware. This distributed approach is presented as the storage solution for Big Data.

12. MapReduce for Big Data Processing

MapReduce is the processing paradigm described by TPointTech for data distributed over a network. It is used to process the distributed data and find the required output.

13. Pig and Hive for Big Data Analysis

After data is stored and processed in the Hadoop ecosystem, TPointTech identifies Pig and Hive as tools that can be used to analyze the data. This separates the analysis stage from the underlying storage and processing stages.

14. Big Data Processing Flow

The TPointTech use case can be summarized as a simple flow:

- Large data is generated from sources such as e-commerce, social networks, weather stations, telecom systems, and stock markets.
- The data is stored in a distributed way using HDFS.
- MapReduce processes the distributed data.
- Pig or Hive can be used for analysis.
- The resulting information can support business decisions such as identifying top customers and understanding buying trends.

This flow combines the sources, issue, and solution described on the TPointTech page.

15. Structured vs Unstructured Data

Type	Description	Example from TPointTech
Structured Data	Data that can be represented in rows and columns.	Bank transaction data
Unstructured Data	Data that is not organized in a traditional tabular structure.	Log files and CCTV footage

TPointTech uses these examples to explain the Variety characteristic of Big Data.

16. Why Big Data Needs Specialized Processing

The central challenge presented by the tutorial is the combination of very large data volumes, different forms of data, and the need to store, process, and analyze that information. The example demonstrates why distributed storage and processing are useful when the dataset is too large for conventional handling.

17. Key Concepts to Remember

- Big Data means data that is very large in size.
- TPointTech uses petabytes (10^15 bytes) as an example scale for Big Data.
- Major sources include social networks, e-commerce, weather stations, telecom companies, and stock markets.
- The 3 V's are Velocity, Variety, and Volume.
- Velocity describes the rapid increase or generation of data.
- Variety describes structured and unstructured forms of data.
- Volume describes the very large amount of data.
- HDFS is used for distributed storage in the presented Hadoop solution.
- MapReduce is used for processing distributed data.
- Pig and Hive can be used for analysis.

18. Conclusion

Big Data is characterized by extremely large quantities of information and can originate from many different sources. The TPointTech tutorial explains the concept through the 3 V's—Velocity, Variety, and Volume—and illustrates how structured and unstructured data can both form part of a Big Data environment.

The e-commerce use case demonstrates the practical challenge of storing, processing, and analyzing huge datasets. TPointTech presents Hadoop-based components such as HDFS for distributed storage, MapReduce for processing, and Pig and Hive for analysis as the solution for the example.